DOCUMENT RESUME

ED 127 345 TM 005 434

AUTHOR Sanders, James R.; Nafziger, Dean H.
TITLE A Basis for Determining the Adequacy of Evaluation Designs.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
SPONS AGENCY Alaska State Dept. of Education, Juneau.
PUB DATE Oct 75
NOTE 57p.; For related documents, see TM 005 430 and 431

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Check Lists; Criteria; Data Collection; Decision Making; *Evaluation; *Evaluation Criteria; *Evaluation Methods; Evaluation Needs; Guidelines; Information Utilization; *Program Evaluation; Program Planning; Standards

ABSTRACT

A basis is provided for judging the adequacy of evaluation plans or evaluation designs in this document. It is assumed that using the procedures suggested to determine the adequacy of evaluation designs in advance of actually conducting evaluations will lead to better evaluation designs, better evaluations, and more useful evaluative information. The paper is divided into four general sections. First, some basic questions are considered--Why evaluate? Why do we need evaluation designs? Why do we need a basis for judging the adequacy of an evaluation design? Answers to these questions serve to underscore the importance of providing a consistent basis for judging evaluation designs. Secodd, a checklist cf basic considerations important in judging evaluation designs is presented. Each component of that checklist is briefly discussed. Third, a sample design is presented, together with an example cf how the checklist can be used in judging an evaluation design. Fourth, noted professional educators' thoughts about judging the adequacy of evaluation designs are presented. This fourth section is intended especially for the reader who would like additional background based upon current literature in the field. (Author/DEP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED127345

A BASIS FOR DETERMINING THE ADEQUACY OF EVALUATION DESIGNS

James R. Sanders

Dean H. Nafziger

Northwest Regional Educational Laboratory

October 1975

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Prepared under contract support from the Alaska Department of Education

TM005 434

A BASIS FOR DETERMINING THE ADEQUACY OF EVALUATION DESIGNS

In recent years, the educational community has widely acknowledged the usefulness of evaluation in providing information about educational programs, policies, and curricula; as a result, evaluation studies are presently an expected--and often mandated--part of most educational programs. At the same time, many evaluation studies fail dismally in their mission of providing helpful and critical decision-making information. Too often such failure is attributable to poor prior planning.

The purpose of this paper is to provide a basis for judging the adequacy of evaluation plans or, as they are commonly called, evaluation designs. The authors assume that using the procedures suggested in this paper to determine the adequacy of evaluation designs in advance of actually conducting evaluations will lead to better evaluation designs, better evaluations, and more useful evaluative information.

To assist the reader, the paper has been divided into four general sections. Readers are encouraged to concentrate on those sections that seem most appropriate for their needs.

First, some basic questions are considered--Why evaluate? Why do we need evaluation designs? Why do we need a basis for judging the adequacy of an evaluation design? Answers to these questions should serve to underscore the importance of providing a consistent basis for judging evaluation designs.

Second, a checklist of basic considerations important in judging evaluation designs is presented. Each component of that checklist is briefly discussed within this section.

Third, a sample design is presented, together with an example of

how the checklist can be used in judging an evaluation design.

Fourth, noted professional educators' thoughts about judging the adequacy of evaluation designs are presented. This fourth section is intended especially for the reader who would like additional background based upon current literature in the field.

We anticipate that the primary audience for this paper will be Alaskan educators and educational administrators--particularly project directors and evaluators--who have to deal with evaluations frequently. The paper is not written for a highly technical audience; the authors recognize that many Alaskan educators--like educators everywhere--have not had time to devote to the detailed study of measurement and statistics. Therefore, in the interest of making the paper useful to the widest possible readership, the criteria presented for judging designs rely on concepts that are easily communicated or commonly known to educators. Technical or otherwise esoteric concepts are deliberately omittted.

Information contained in this paper can be used in two ways. First, it can be used by evaluators as a guide in preparing--and later reviewing and improving--their own evaluation designs. Second, project directors can use the checklist to judge the adequacy of evaluation designs submitted to them. Special communication needs often arise between an evaluator and project director; evaluation designs can facilitate clear communication, and serve as a standard to assure quality evaluation. An evaluation design provides a written record of decisions about the evaluation to which both the evaluator and project director can refer.

I. BASIC QUESTIONS REGARDING EVALUATION

As a preface to the checklist of criteria for determining the adequacy of evaluation designs, a few basic questions relating to evaluation are briefly addressed in this section. Answers to these questions amplify the assumptions and rationale underlying this paper.[1]

Why Evaluate?

Evaluation gives information about the quality of educational programs provided to our children. Without it, we could not know whether a curriculum was effective, whether a student was performing satisfactorily, or whether the dollars earmarked for education were being spent well.

Given the benefits it provides, proper evaluation is an essential part of all education. Those benefits may include the following:

1. Identification of strengths and weaknessess--a first step toward improvement.
2. Detection of problems before correction becomes difficult or impossible.
3. Identification of needs that should be addressed through educational action.
4. Identification of human and other resources that can be used effectively in education.

---

[1] To meet the anticipated needs of the audience for this paper, discussion of the questions is abbreviated. For a more complete explication of some of these questions and others (e.g., When should evaluation be done? When should an external rather than an internal evaluation be used?), see Wright, W.J., & Worthen, B.R., Standards and Procedures for Development and Implementation of an Evaluation Contract. A discussion paper prepared for the Alaska Department of Education, October 15, 1975.

5. Documentation of desired outcomes of education.
6. Information useful in educational planning and decision making.
7. Cost information that can ultimately reduce educational expense.

Why Do We Need Evaluation Designs?

Everyone implicitly engages in evaluation virtually every day of his life. When buying a new coat or choosing a restaurant we make decisions based on our evaluation of the quality of the available choices. These evaluations are often informal and are seldom planned in terms of procedures and outcomes. Given time constraints and the relative low penalties for making errors, such informal evaluations are entirely appropriate. However, when the choices or courses of action affect students, result in expenditures of scarce public funds, or involve long term commitments or benefits, the situation is different.

Carefully planned evaluation procedures, which are referred to in this document as designs, help both the project director and the evaluator understand the process through which a program or project will be judged. The design also provides for the organization of resources and activities which are required for an evaluation study.

Preparation and use of an evaluation design has benefits for both the evaluator and the project director. Presenting an evaluation design gives the evaluator an opportunity to communicate with project staff concerning proposed evaluation procedures and ensure their clear understanding of the process. At this point changes can be made without disrupting the evaluation. For the project director and staff, an evaluation design provides an opportunity to review the type of information to be

yielded by the evaluation so that additional or alternative types of data collection can be suggested if necessary to provide complete information to all users of the evaluation results. Also, evaluation procedures can be reviewed in order to ensure that no unexpected disruptions of the program will occur. Many misunderstandings have occurred between evaluator and project staff, and many an evaluation has altered in focus because a clear, systematic evaluation design was not prepared early in the evaluation.

The advantages to completing an evaluation design early include the following:

1. Assuring clear and accurate direction for the study by establishing the uses for evaluation results, and by specifying expected products of the evaluation.
2. Assuring completeness of procedures by giving others an opportunity to make suggestions.
3. Identifying inconsistencies in perceptions by the evaluator and project director of evaluation plans so that these can be resolved prior to actual evaluation.
4. Providing a clearly defined set of tasks for the evaluation so that attention is maintained on important outcomes.
5. Assuring efficiency in the evaluation by organizing resources and activities. (Like any substantial educational undertaking, evaluation requires good management and accounting.)

In short, evaluation design helps the evaluator and project director communicate clearly about the project. Because of the importance of the design, it is critical that it be closely scrutinized and all details dis-

cussed. Specific criteria or guidelines are particularly helpful to clients in critically reviewing a design.

Why Do We Need A Basis for Judging the Adequacy of an Evaluation Design?

Most school administrators have few, if any, persons on their staffs with sufficient training and experience in evaluation to judge the adequacy of evaluation designs solely on the basis of their own knowledge. In addition, qualified persons are in such demand that they are often unable to spend the time necessary to personally review all evaluation designs used in the system. Therefore, administrators and other educators are often left with little or no help in determining whether designs proposed for evaluations of their programs are sound and capable of providing useful information about those programs. Given this situation, there is a need for written guidelines which might serve as a basis for judging an evaluation design. Several benefits are expected to accrue from the use of such guidelines:

1. The guidelines should improve the quality of evaluation. Established guidelines should represent what is known about producing useful, technically correct evaluations, and their use should therefore preclude many errors common to evaluation studies.
2. The guidelines should provide a framework for developing evaluation designs. Established guidelines clarify and make public the expectations about what a good evaluation design ought to include. Because they aid communication in this way, guidelines can be used as a basis for designing evaluations.

3. The guidelines should assist administrators in monitoring evaluation work. The use of guidelines ensures that important aspects of an evaluation will be described in the design, and that descriptions will be specific enough to assist in monitoring the evaluation study.
4. The guidelines can help address ethical considerations in contract evaluation work. Established guidelines help guarantee that aspects of the evaluation which are subject to questions of ethics--such as reporting procedures, information release and dissemination policies--will be considered, and relevant issues resolved prior to the evaluation study. This in turn helps prevent inappropriate use of the evaluation results.

Ethical conduct in educational evaluation is a critical issue which pervades much of the current literature on evaluation. Unfortunately the scope of this paper does not permit an adequate discussion of the topic. A comprehensive treatment of ethical standards and conduct, while in order must await another document devoted specifically to that issue.

II. A CHECKLIST FOR JUDGING THE ADEQUACY OF EVALUATION DESIGNS

Virtually everyone involved in any way with evaluation is concerned with the quality of the evaluation effort. The checklist presented on the following pages provides a basis for judging the adequacy of evaluation designs. The checklist is divided into four general sections, each of which covers several criteria regarding evaluation designs. Those criteria are addressed through a set of related questions. All criteria are more thoroughly discussed following the presentation of the checklist.

Briefly, the four general sections are as follows. The first section includes Criteria concerning the adequacy of evaluation planning which covers such issues as whether the proposed evaluation addresses all important aspects of the program, and whether the evaluation can be completed within existing constraints.

The second section includes Criteria concerning the adequacy of the collection of and processing of information. These questions cover the reliability, objectivity, and representativeness of the information obtained.

The third section, Criteria concerning the adequacy of the presentation and reporting of information, deals with the usefulness and completeness of the anticipated reports.

The fourth section includes General Criteria, those which deal with ethical considerations and protocol.

CHECKLIST FOR JUDGING THE ADEQUACY OF AN EVALUATION DESIGN

Directions: For each question below, circle whether the evaluation design has clearly met the criterion (Yes), has clearly not met the criterion (No), or cannot be clearly determined (?). Circle NA if the criterion does not apply to the evaluation design being reviewed. Use the Elaboration column to provide further explanation for criterion where a No or a ? has been circled.

Title of Evaluation Document: _____

Name of Reviewer: _____

Criterion	Criterion Met	Elaboration
I. Regarding the Adequacy of the Evaluation Conceptualization		
A. Scope: Does the range of information to be provided include all the significant aspects of the program or product being evaluated?		
1. Is a description of the program or product presented (e.g., philosophy, content, objectives, procedures, setting)?	Yes No ? NA	
2. Are the intended outcomes of the program or product specified, and does the evaluation address them?	Yes No ? NA	
3. Are any likely unintended effects from the program or product considered?	Yes No ? NA	
4. Is cost information about the program or product included?	Yes No ? NA	

Criterion	Criterion Met	Elaboration
B. Relevance: Does the information to be provided adequately serve the evaluation needs of the intended audiences?		
1. Are the audiences for the evaluation identified?	Yes No ? NA	
2. Are the objectives of the evaluation explained?	Yes No ? NA	
3. Are the objectives of the evaluation congruent with the information needs of the intended audiences?	Yes No ? NA	
4. Does the information to be provided allow necessary decisions about the program or product to be made?	Yes No ? NA	
C. Flexibility: Does the evaluation study allow for new information needs to be met as they arise?		
1. Can the design be adapted easily to accommodate new needs?	Yes No ? NA	
2. Are known constraints on the evaluation discussed.	Yes No ? NA	
3. Can useful information be obtained in the face of unforeseen constraints, e.g., noncooperation of control groups?	Yes No ? NA	

Criterion	Criterion Met	Elaboration
D. Feasibility: Can the evaluation be carried out as planned?		
1. Are the evaluation resources (time, money and manpower) adequate to carry out the projected activities?	Yes No ? NA	
2. Are management plans specified for conducting evaluation?	Yes No ? NA	
3. Has adequate planning been done to support the feasibility of particularly difficult activities?	Yes No ? NA	
II. Criteria Concerning the Adequacy of the Collection and Processing of Information		
A. Reliability: Is the information to be collected in a manner such that findings are replicable?		
1. Are data collection procedures described well enough to be followed by others?	Yes No ? NA	
2. Are scoring or coding procedures objective?	Yes No ? NA	
3. Are the evaluation instruments reliable?	Yes No ? NA	

Criterion	Criterion Met	Elaboration
B. Objectivity: Have attempts been made to control for bias in data collection and processing? 1. Are sources of information clearly specified. 2. Are possible biases on the part of data collectors adequately controlled?	Yes No ? NA Yes No ? NA	
C. Representativeness: Do the information collection and processing procedures ensure that the results accurately portray the program or product? 1. Are the data collection instruments valid? 2. Are the data collection instruments appropriate for the purposes of this evaluation? 3. Does the evaluation design adequately address the questions it was intended to answer?	Yes No ? NA Yes No ? NA Yes No ? NA	
III. Criteria Concerning the Adequacy of the Presentation and Reporting of Information		
A. Timeliness: Is the information provided timely enough to be of use to the audiences for the evaluation? 1. Does the time schedule for reporting meet the needs of the audiences? 2. Is the reporting schedule shown to be appropriate for the schedule of decisions?	Yes No ? NA Yes No ? NA	

Criterion	Criterion Met	Elaboration
B. Pervasiveness: Is information to be provided to all who need it. 1. Is information to be dissemirated to all intended audiences? 2. Are attempts being made to make the evaluation information available to relevant audiences beyond those directly affected by the evaluation.	Yes No ? NA Yes No ? NA	
IV. General Criteria		
A. Ethical Considerations: Does the intended evaluation study strictly follow accepted ethical standards? 1. Do test administration procedures follow professional standards of ethics? 2. Have protection of human subjects guidelines been followed? 3. Has confidentiality of data been guaranteed?	Yes No ? NA Yes No ? NA Yes No ? NA	
B. Protocol: Are appropriate protocol steps planned? 1. Are appropriate persons contacted in the appropriate sequence? 2. Are Department policies and procedures to be followed.	Yes No ? NA Yes No ? NA	

Use of the Checklist

The checklist should be used like any other set of guidelines. Once the design has been read thoroughly, each item on the checklist should be considered with respect to the design. For each question related to the criteria, one of the four available options--Yes, No, ?, Not Applicable (NA)--should be circled, depending on whether the criterion was adequately met.

Each question should be clearly and fully addressed by the evaluation design. If that is the case and if the requirements of the question are met, the reviewer should circle "Yes." For any question which is not discussed or the requirements of the question not met, the reviewer should circle "No." If for some reason--such as inadequate information--it cannot be determined whether the question is appropriately answered, the reviewer should circle "?." If a question is not applicable to a particular evaluation, the reviewer should circle "NA."

In the space marked "Elaboration" the reviewer should note any additional comments that ought to be transmitted to the author of the evaluation design. In particular, if a criterion was not met or if there was some question about its being met, elaboration would be warranted. Further, ambiguous intentions or plans seeming to require revision should all be noted in the "Elaboration" section. Upon completing the checklist, it should be given to the evaluator and to others affected by the evaluation so that it can be used to revise the evaluation design.

There will likely be instances in which the reviewer will want to obtain advice from another person about whether a question has been appropriately answered. For example, this might occur when judging information about the validity of a test or about the appropriateness of a data collection design. The user of the checklist should always seek

and obtain advice when the content of an evaluation design or items on the checklist prevent him from making a judgement.

It is important to remember that an evaluation design is a vehicle for communication between an evaluator and those whose role calls for reviewing the evaluation plan. The checklist helps organize that communication. In cases where an evaluation is conducted by a contractor, the design becomes a vehicle for a communicator between the evaluator and the client. In such cases the checklist assists a client in judging adequacy of the design, and provides a basis for giving feedback to the evaluator. If the evaluator is involved in the program being evaluated, the guidelines provide a basis for the evaluator and his or her colleagues to check the design.

Each major point of consideration noted in the checklist is reviewed in the next few pages, along with information that should be covered in evaluation design.

CRITERIA CONCERNING THE ADEQUACY OF EVALUATION PLANNING

A. Scope. The evaluation design should include plans to collect information about all significant aspects of the program, product, or process being evaluated. If a student's performance is being evaluated, and the evaluation design does not call for collecting information about conditions that might adversely affect his or her performance, that oversight should be noted. The primary concern within this criterion is whether the focus of the evaluator's attention is too narrow.

B. Relevance. The design should include plans to collect information that addresses the concerns of those who requested the evaluation. For example, if a compensatory education project

is being evaluated and the project director is concerned about upgrading the reading skills of children in the program, the evaluation design should call for collecting information about improvement in children's reading skills. To make the design relevant to the needs of the evaluation audiences, the evaluator should indicate the various audiences that need information and give the expected uses of the information. Any suggestions or changes concerning the information to be collected should be noted.

C. Flexibility. The evaluation design should be open enough to allow for the addition of new information gathering and processing activities. This is especially important in complex, long term program evaluations where changes in program plans are likely. If a new program directed toward changing the attitudes of minority children toward school is just getting underway and the evaluation design does not allow for changes in instrumentation resulting from changes in program objectives, it should be noted that the criterion is not met, and suggested means of allowing for such change should be given.

D. Feasibility. The evaluation design should provide enough information so that the feasibility of carrying out the study can be determined. Many evaluation designs fail to meet this criterion. Feasibility can be determined on the basis of schedules, budget, personnel assigned to conduct specific activities, proposed procedures in data collection, and reporting plans. An evaluation design is not useful unless it can actually be implemented.

CRITERIA CONCERNING THE ADEQUACY OF THE COLLECTION AND PROCESSING OF INFORMATION

A. Replicability. The evaluation design should include procedures for assuring that the information being collected is accurate and that if the evaluation were replicated the same results would occur. Statistical reliability indices should be provided for standardized instruments, and procedures for determining the reliability of information collected by nonstandardized instruments should be included in the evaluation design. The reviewer should check the design to see whether such information is provided and circle the appropriate response. If the design provides no way to check the accuracy or replicability of information being collected, those concerns should be described.

B. Objectivity. The evaluation design should incorporate procedures to control for biases. Those biases that may affect an evaluator's collection or interpretation of information should be clearly labeled and minimized. Methods for maintaining fairness and object vity--such as the use of external data collectors, objective and unbiased instrumentation, or interpretation panels for reporting findings--should be incorporated into an evaluation design whenever possible. If the reviewer has concerns about inherent bias in the evaluation design, those concerns should be noted and discussed with the evaluator.

C. Representativeness. The information to be collected should accurately represent the program or project being evaluated. Data collection instruments should be valid, and they should obtain information that bear upon all the evaluation questions. Information about all significant aspects of the program should be reported. Sampling pro-

cedures are often used when the amount of information needed for a complete picture becomes too unweildy. When this is done, representative samples should be selected.

CRITERIA CONCERNING THE ADEQUACY OF THE PRESENTATION AND REPORTING OF INFORMATION

A. Timeliness. The evaluation design should describe how reports and other presentations fit into the schedule for decision making. Report deadlines should reflect the informational needs of the persons to whom the presentations are directed. The design should contain a reporting schedule and content descriptions of reports or other presentations, and show the relationship to the decision-making schedule.

B. Pervasiveness. The evaluation design should call for the delivery of reports or presentations to all relevant audiences. These include any persons or groups that affect or are affected by the evaluation itself or the object of the evaluation. Suggestions about the distribution of evaluation information should be recorded under "Elaboration."

GENERAL CRITERIA

A. Ethical Considerations. The evaluation design should cover whatever ethical considerations may be of concern. In some cases certain information obtained through the evaluation may be confidential, and steps to protect confidentiality should be included in the design. An evaluator should also be aware that some data collection procedures--such as use of peer informers--may be threatening to subjects and such practices should be avoided. Additional ethical considerations not addressed within the design should be noted under "Elaboration."

B. Protocol. The evaluation design should include some consideration of protocol. For example, it is often necessary to obtain a superintendent's permission to talk to a building principal or teacher before actually contacting that person. In many cases, it is professional courtesy to request permission to use the work of others before referencing it. In all phases of information collection and reporting, strict protocol should be observed.

Summarizing the Information Contained in the Checklist

After considering each question on the checklist, a reviewer will have a series of circled responses in one column and a number of comments in the other. "No" or "?" responses indicate a need for additional information. Comments in the "Elaboration" section will provide a basis for making various sorts of improvements in the design. In short, the information from the checklist summarizes for the evaluator what changes are needed to make the evaluation design acceptable.

Whenever evaluation is conducted under contract, the evaluation design becomes an important focus of communication among the evaluator, his staff, and the client. Modifying the design to make it acceptable to both sides can aid that communication process. Should irreconcilable differences arise between evaluator and client, one alternative is to terminate the relationship; another is to bring in an objective outsider to negotiate changes. In most cases, however, differences can be resolved through design modification.

The following section of the paper provides a sample application of the checklist; that sample application is intended to clarify concepts described in this section. The reader is encouraged to gain experience

in using the checklist by first applying it to the design, and then comparing his results with those of the authors.

III. EXAMPLE APPLICATION OF THE CHECKLIST TO AN EVALUATION DESIGN

The checklist for judging evaluation designs that is given in the previous section is to be used as a tool to help identify strengths and weaknesses in an evaluation design. Identified weaknesses can then be improved before the evaluation begins.

In this section, the checklist is applied to a fictitious evaluation design. There are two parts to this section of the paper. The first is a short, fictitious evaluation design. This design is not intended to represent any actual evaluation study in Alaska or elsewhere. Any resemblance to an existing evaluation study in Alaska is purely coincidental. Rather, the design represents the type of evaluation designs frequently encountered by project directors and other administrators. The design is neither all good nor all bad. As will be seen, it contains some components that are entirely adequate and others that require improvement.

The second part of this section of the paper is the actual application of the checklist. Each question in the checklist is answered for the fictitious design, and an explanation of each answer is given.

EVALUATION DESIGN FOR THE HARTMAN READING PROGRAM
FOR FIVE BOROUGHS

Introduction

In recent years reading instruction has become a major target area for education not only in Alaska but throughout the United States. As a result of this emphasis, several new reading programs, textbooks, and instructional materials have been developed.

Recently, one of these new programs, the Hartman Reading Program, was adopted jointly by five Alaskan boroughs: Elk Mountain, Donelly, Banks, Karnaska, and Port. The Hartman Reading Program is appropriate for students in grades one through six. It was selected because it had been developed for use in a variety of cultural settings, and because it purported to improve the self-concept of students from minority cultural groups. The expense involved in adopting the Hartman Reading Program was too much to be borne by any one borough alone, but a joint effort made adoption feasible.

The purpose of this evaluation is to determine whether the Hartman Program is fulfilling the goals which the five boroughs have set for new reading programs.

Program Goals and Evaluation Questions

The five-borough Planning Committee which selected the Hartman Reading Program have established four goals that any new reading program within those boroughs is expected to attain. These four goals are listed below along with several associated evaluation questions.

Goal 1: Children in the program will achieve in all reading subjects at a rate commensurate with their own age, ability, and grade level.

Question 1.1: How does the performance of children in the new program, as measured on a standard reading achievement test, compare to that of other children in the United States at the same grade level?

Question 1.2: How does the performance of children in the new program, as measured on standard reading achievement tests compare to the performance of children in the district in past years?

Question 1.3: How does the performance of children in the new program compare to that of children in the old reading program?

Goal 2: Children in the new program will demonstrate growth in self-esteem and improvement in self-concept.

Question 2.1: How do children in the new program compare with children in the old program in measures of self-esteem and self-concept?

Goal 3: All teachers and staff members of participating classrooms will be involved in a comprehensive inservice training program.

Question 3.1: What percentage of teachers and staff members from participating classrooms have taken the voluntary training program?

Question 3.2: To what extent do teachers and staff members express satisfaction with the training program?

Goal 4: Parents will be involved in the implementation of the new program.

Question 4.1: What percentage of parents of students in participating classrooms become involved in the classroom activities designed for parents?

Audiences for the Evaluation

The primary audience for the evaluation is the Planning Committee for the five boroughs. Based upon the results of the evaluation, the Planning Committee will decide to adopt the Hartman Reading Program throughout the five boroughs, or to eliminate use of the program. That decision will be made in July.

One secondary audience for the evaluation is teachers throughout the boroughs. Data collected during the pretest can be used by teachers to diagnose reading difficulties and poor self-concepts by students.

Another secondary audience consists of project directors, evaluators, and other educators throughout the state who would like information about the Hartman Reading Program or about the evaluation procedures used in this study.

Data Collection Design for the Hartman Reading Program

In order to allow for classroom differences while making necessary comparisons, a pre-post-test, treatment-control group design was developed. Students in the new program are designated the treatment or experimental group, and those in the regular school program are considered the control group. Three alternative methods for gathering comparative data have been designed. Each of these designs depends on random assignment of students or classrooms to treatment and comparison groups at the beginning of the school year. The alternatives are listed below in order of

desirability. Since more desirable designs may also be more difficult to implement, the most desirable alternative that can be implemented within the constraints imposed by the school situation will be chosen.

Alternative I: Random Assignment of Students Within Classrooms

This experimental design allows for random assignment of students to program and control groups within classrooms. This design is based on the assumption that such assignments are acceptable to teachers, and that the two reading programs can be implemented in each classroom.

Student Selection Procedure.

1. Determine, by grade and classroom, the number of students who would participate in the program.
2. Make an alphabetical list, by classroom, of students who may be selected to fill program to capacity. (This list should contain twice the number of students needed to fill program quota.)
3. Alternately assign names to program and control groups in each classroom, as follows: first name on list to program; second name to comparison group; third name to program; fourth name to comparison, etc.

Alternative II: Random Assignment of Classes

The second alternative involves the random assignment of entire classes to treatment and control groups. It assumes that several classes of students at each grade level can adopt the new program or remain with the old one.

Classroom Selection Procedure.

1. Determine, by grade, the number of students who would participate in the new program.

2. Prepare a list, by grade, of classes which would participate in the program. Assign a number to each classroom on the list.
3. Use a random number table to select classes to participate in the treatment group, and choose half of the classes for that purpose. The remainder will constitute the control group.

Alternative III: Teacher Selection of Program

This alternative allows teachers to choose whether they would like to participate in the new program or keep using the old one. The selection procedure simply involves allowing teachers to choose according to their preferences.

The comparison design will be used to determine the effects of the Hartman Reading Program in the areas of reading performance and self-concept. Statistical techniques appropriate for the design chosen will be used. Comparative analysis of differences in performance on all pre-post-tests will be included in the design. The specific question answered here is whether children in the program are learning significantly more than comparable children not in the program.

Reporting Procedures

Three types of reports will be prepared--a Teacher Report for each teacher, an Administrative Report, and a Technical Report.

A Teacher Report will be compiled for each teacher's classroom summarizing pretest data for the classroom. The teacher feedback report will include:

. Tables (two per class) showing scores, percentiles, and stanines for each pupil on each test.

- Tables (two per class) of profiles showing graphically the percentile equivalents of the average score for each test and comparison of each child with his class, with children in other classes, and with students at the same grade level in other schools tested.

- Local norms and standardization as given in administrative feedback report.

- An interpretive guide for using the data provided.

The Administrative Report will include a summary of the comparison study results. The effects of the Hartman Reading Program in comparison with the standard program will be summarized and interpretations given.

The Technical Report will include:

- Detailed description of data-collecting methods and procedures.
- Detailed description of procedures used in data analysis throughout the project.
- Summary tables as presented in administrative feedback.
- Item analysis of all tests used in project.
- Norms on all tests used in project.

The Administrative Report and Technical Report will be reviewed by a panel of teachers, administrators, State Department of Education personnel and university educators to determine the accuracy, fairness, and impartiality of the reports. Reports will be revised on the basis of those reviews, and, if consensus is not reached, an addendum giving the opposing interpretations will be attached.

Description of Program and Comparison Treatment

Program groups will receive reading instruction as described in the Hartman Reading Program Guide for Instruction. The Guide gives a detailed account of materials to be used, involvement of parents, sequencing of concepts, and time required for each activity. The Guide also provides the philosophical underpinning of the program, general program objectives, and settings in which the program should be used. Because the Guide is readily available, the program description is not repeated in this design. The comparison group will receive instruction in the usual curriculum offered in the five boroughs. Because the same curriculum is used in each of the boroughs, no further standardization of treatment will be required. A detailed description of the standard curriculum and its implementation is provided in the Curriculum Guide.

Testing Instruments

Tests were chosen to measure important reading skills being taught in the reading programs of the boroughs. These skills encompass listening and writing as well as more typical reading skills. In addition, a test of self-esteem is included. The tests chosen--the Sequential Tests of Educational Progress, the Multicultural Reading Series, and the Self-Observation Scale--are described on the following pages.

The Sequential Tests of Educational Progress (STEP) are achievement-oriented tests. These instruments measure the broad outcomes of general education, focusing on the ability to solve new problems on the basis of

information learned as opposed to ability to handle only "lesson material." The STEP instruments provide for continuous measurement of skills over nearly all of the years of general education; therefore, they measure more of the cumulative effect of instruction.

The STEP Listening tests were designed to measure a student's ability to understand, interpret, apply and evaluate what he listens to. The listening skills are broken down into sub-abilities which are classified as follows: plain-sense comprehension, interpretation, evaluation and application.

The STEP Listening tests include typical examples of what might actually be said to students in a school situation. Each test includes materials of the following types: direct and simple explanation, exposition, narration, argument and persuasion, and aesthetic material (both poetry and prose).

These tests are available for grade four to college sophomore level. They are subdivided into four levels of difficulty to provide for a wide range of abilities.

STEP Listening test interpretation begins with a score which is translated into percentiles through the use of normed tables. The publisher also provides national norms from a sample of students' scores with those of a nationwide sample of students at the same educational level. Directions for constructing local STEP norms are provided.

The STEP Writing test measures ability to think critically in writing, organizing materials, choosing appropriate materials to write effectively, and using appropriate, conventional punctuation and grammar.

The materials chosen were those from actual student writing excerpted from letters, newspapers, answers to test questions, reports, stories,

notes, outlines, questionnaires and directions.

The STEP Writing test is based on the same criteria as the listening test. Norms were formulated in the manner described in the listening section.

The tests of reading in the Multicultural Reading Series are designed to measure both vocabulary and comprehension. At grade levels beyond primary one, comprehension is measured by two subtests: speed of comprehension, and level of comprehension.

Scores on the tests of reading may be used not only as measures of achievement in reading itself, but also as bases for estimating ability to achieve. In grouping children and adjusting instruction to individual differences, a measure of reading ability is often useful as a measure of mental ability. After a child has learned to read, the use of both measures is much better than the use of either one alone.

The test was constructed by the Testing Research Associates (1962) especially for multicultural student populations. Administration time varies from 30 to 50 minutes. Given specific instructions, a teacher may administer the test successfully.

The technical report of the series presents an average parallel test reliability of .87 and an average correlation of .78 with the STEP; this indicates a relatively high concurrent validity.

The Self-Observation Scales (SOS) is a direct, self-report, group-administered instrument comprising 45 items (Forms A and B) designed to measure five dimensions of children's affective behavior: self-acceptance, social maturity, school affiliation, self-security and achievement motivation. The SOS has been translated into various languages including Spanish, Italian, Chinese, Greek, Korean, Japanese, Tagalog

and Arabic.

The Technical Bulletin (No. 1) for the SOS reports the following split-half reliability values (N=4144):

	Self-Acceptance	Social Maturity	School Affiliation	Self-Security	Achievement Motivation
Form A	.75	.77	.76	.81	Not Available (NA)
Form B	.79	.79	.79	.81	NA

Intersubscale correlations are reported as follows (N=4144):

	Self-Acceptance	Social Maturity	School Affiliation	Self Security	Achievement Motivation
Self-Acceptance	-	.06	.48	.18	NA
Social Maturity	-	-	.34	.58	NA
School Affiliation	-	-	-	.36	NA
Self-Security	-	-	-	-	NA

Content validity is assured by publishers at the Institute for Development of Educational Auditing.

The validation and norming sample includes students from 150 schools nationwide. In drawing the sample, particular attention was paid to the social, geographic, and socioeconomic characteristics of the participating schools. The norm group was composed of 9,030 students at K - 3 levels.

The validation and norming sample includes students from 150 schools. The norm group was composed of 9,030 students at K-3 levels.

According to the publishers, "The SOS differs from other similar instruments in (a) the extensive validation study which has accompanied the national norming effort, (b) the emphasis on the healthy and positive, rather than pathological and negative dimensions of children's affective behavior, and (c) the practical decision-making orientation rather than a research, theoretical orientation."

Other Data Collection Forms

Data about the participation of teachers and staff members in inservice programs will be collected from the records of inservice instructors. The satisfaction of teachers with the training will be measured using the Training Satisfaction Questionaire (TSQ). The TSQ has been used frequently in the boroughs. It consists of 20 questions about the training, and has adequate reliability (KR-20 coefficient = .83) for this type of questionnaire.

Participation of parents in classroom activities will be determined using a form to be filled out by teachers and a questionnaire to be sent to parents. Information from these two instruments will be cross checked and discrepancies resolved by the evaluation team with follow-up correspondence.

Procedure Clearance Steps

All data collection activities, teacher training workshops, evaluation questionnaires, and mass communication strategies will be submitted to the chief school officer in each borough for approval prior to use. Procedures for implementing any evaluation plans will be determined jointly with the chief school officer.

Evaluation Activities Time Line

September

Select treatment and control groups

Request student names and identification numbers

Deliver test materials to schools

Conduct pretest evaluation inservice

October

Submit completed student I.D. blanks to evaluation unit

Administer pretests

Pick up completed pretests from schools

Visit schools evaluation team

November

Administer listening tests

Mail student information blank to schools

Complete and deliver individual Teacher Reports

December

Begin class observation schedule

Submit completed student information blank to evaluation unit

Classroom observation schedule (ongoing)

January

Monitor experimental/comparison groups and continue classroom observations

Conduct evaluation conference for parents/advisory council members

Classroom observation (ongoing)

February

Participate in visits to schools

Classroom observation (ongoing)

March

Continue classroom observations and monitoring of experimental/comparison groups

Continue participation in visits to schools

Classroom observation (ongoing)

April

Mail parent/teacher/administrator questionnaires

Conduct posttest inservice

Classroom observation (ongoing)

Questionnaires due in the evaluation unit by the end of the month

May

Deliver posttest materials to schools

Posttest administration

Completed posttests to be picked up

June

Technical Report and Administrative Report completed

July

Use of reports for adoption or elimination of the use of the Hartman Reading Program

Use of the Checklist with the Fictitious Evaluation Design

In this section the fictitious evaluation design is reviewed to demonstrate the use of the checklist in determining the adequacy of an evaluation design. The rationale for each response is provided immediately following each set of questions on the checklist. These elaborations are somewhat longer than would be provided by most users of the checklist.

I. Regarding the Adequacy of the Evaluation Conceptualization	
A. Scope: Does the range of information to be provided include all the significant aspects of the program or product being evaluated?	
1. Is a description of the program or product presented (e.g., philosophy, content, objectives, procedures, setting)?	(Yes) No ? NA
2. Are the intended outcomes of the program or product specified, and does the evaluation address them?	(Yes) No ? NA
3. Are any likely unintended effects from the program or product considered?	Yes (No) ? NA
4. Is cost information about the program or product included?	Yes (No) ? NA

The criterion of Scope seems to be only partially met in the design. The first two of the four questions can be answered with a "yes." The design does include a description of the Hartman Reading Program, although it is done by referencing the Guide for Instruction for the program. (see page 7)[2].

[2]Note that here, as will always occur with the use of any checklist, the user's professional judgment must guide decisions about how well questions have been answered and criteria met. Some users may wish the program description from the Guide to be included in the design as an appendix or in the test itself before a "Yes" is circled. This is certainly justified. The important point is that provision be made to give an adequate description of the program to those who need it.

Also, the objectives of the program are given through a series of questions that relate the general goals of the planning Committee (see pages 1 and 2).

On the last two questions the evaluation design does not fare as well. No provision is made for any unintended effects that might occur from the use of the program. Neither is any information given about the cost of the program. In order for the criterion of Scope to be adequately met, the two types of missing information should be included.

B. Relevance: Does the information to be provided adequately serve the evaluation needs of the intended audiences?				
1. Are the audiences for the evaluation identified?	(Yes)	No	?	NA
2. Are the objectives of the evaluation explained?	(Yes)	No	?	NA
3. Are the objectives of the evaluation congruent with the information needs of the intended audiences?	(Yes)	No	?	NA
4. Does the information to be provided allow necessary decisions about the program or product to be made?	(Yes)	No	?	NA

The evaluation design has adequately met the criterion of Relevance. Primary and secondary audiences were identified (page 3). The objectives of the evaluation were delineated in a set of questions that followed from the information needs of the primary audience. Further, decisions about the program can be made on the basis of the answers to the evaluation questions.

C. Flexibility: Does the evaluation study allow for new information needs to be met as they arise?				
1. Can the design be adapted easily to accommodate new needs?	(Yes)	No	?	NA
2. Are known constraints on the evaluation discussed.	Yes	(No)	?	NA
3. Can useful information be obtained in the face of unforeseen constraints, e.g., noncooperation of control groups?	(Yes)	No	?	NA

The evaluation design seems to be reasonably successful regarding the criterion of Flexibility. It seems that the proposed evaluation would be able to accomodate new information needs because several data collection procedures and instruments are to be employed. In general, an evaluation that uses several procedures is more flexible than an evaluation that relies heavily on one or two methods or instruments. Another strength of the design is that there is a set of alternatives for gathering comparative data. Selection of groups for a comparison study is typically an area in which some flexibility is needed.

A weakness regarding the Flexibility criterion is that there is no discussion of the constraints on the study. Nearly all evaluation studies are subject to constraints of various degrees of importance, and they should be explained in the design.

D. Feasibility: Can the evaluation be carried out as planned?				
1. Are the evaluation resources (time, money and manpower) adequate to carry out the projected activities?	Yes	No	(?)	NA
2. Are management plans specified for conducting evaluation?	Yes	(No)	?	NA
3. Has adequate planning been done to support the feasibility of particularly difficult activities?	Yes	No	(?)	NA

The adequacy of the evaluation design as it relates to the Feasibility criterion is in question. The available resources to conduct the study are not given, and so no judgment can be made about their adequacy. There is no management plan which lists the major tasks, time required to complete tasks, or personnel. Also, there is only a

little evidence that particularly difficult tasks are feasible. Clearly, more information relating to the feasibility of the study is needed.

II. Criteria Concerning the Adequacy of the Collection and Processing of Information	
A. Reliability: Is the information to be collected in a manner such that findings are replicable?	
1. Are data collection procedures described well enough to be followed by others?	(Yes) No ? NA
2. Are scoring or coding procedures objective?	(Yes) No ? NA
3. Are the evaluation instruments reliable?	(Yes) No ? NA

Adequate information supporting the Replicability criterion seems to be included. The tests and questionnaires to be used in the study are described in adequate detail, and their reliability is shown to be sufficiently high (pages 7ff). In the one instance where low reliability of data may occur--teacher and parent reports of parent involvement--the data are to be cross checked (page 11).

B. Objectivity: Have attempts been made to control for bias in data collection and processing?	
1. Are sources of information clearly specified.	(Yes) No ? NA
2. Are possible biases on the part of data collectors adequately controlled?	Yes No ? (NA)

The Objectivity criterion seems to have been met. It is clear from whom each type of data will be collected. Further, there do not seem to be any particular threats to the objectivity of the data, and so no special controls are required. Hence, the "NA" for the second question.

C. Representativeness: Do the information collection and processing procedures ensure that the results accurately portray the program or product?	
1. Are the data collection instruments valid?	Yes No (?) NA
2. Are the data collection instruments appropriate for the purposes of this evaluation?	Yes (No) ? NA
3. Does the evaluation design adequately address the questions it was intended to answer?	(Yes) No ? NA

The Representativeness criterion has not been met satisfactorily in this design. The inadequacies with respect to this criterion are brought to light by the first two questions. First, the validity of the achievement tests is open to question. No information about the validity of the Sequential Test of Educational Progress is provided, although such information may well be available. Some Validity information is given for the Multicultural Reading Series (page 9). For the Self Observation Scale, only an ambiguous statement about validity is given (page 10).

III. Criteria Concerning the Adequacy of the Presentation and Reporting of Information	
A. Timeliness: Is the information provided timely enough to be of use to the audiences for the evaluation?	
1. Does the time schedule for reporting meet the needs of the audiences?	(Yes) No ? NA
2. Is the reporting schedule shown to be appropriate for the schedule of decisions?	(Yes) No ? NA

The evaluation design clearly meets the criterion of Timeliness. The needs of the audiences were taken into account and a reporting schedule was developed consistent with those needs (page 13).

B. Pervasiveness: Is information to be provided to all who need it.				
1. Is information to be disseminated to all intended audiences?	(Yes)	No	?	NA
2. Are attempts being made to make the evaluation information available to relevant audiences beyond those directly affected by the evaluation.	Yes	No	(?)	NA

The Pervasiveness criterion is met partly in that the intended audiences for the evaluation are to receive adequate information. However, there are possible unintended audiences that have been largely ignored. The only report to be made available on a broad scale is the Technical Report. Other people who might benefit from information from the evaluation should be considered, and an appropriate report should be written for them. For example, a general summary of the major effects of the Hartman Reading Program would probably be useful information for many superintendents and principals to have.

IV. General Criteria				
A. Ethical Considerations: Does the intended evaluation study strictly follow accepted ethical standards?				
1. Do test administration procedures follow professional standards of ethics?	(Yes)	No	?	NA
2. Have protection of human subjects guidelines been followed?	Yes	(No)	?	NA
3. Has confidentiality of data been guaranteed?	Yes	(No)	?	NA

The criterion of Ethical Considerations does not seem to have been completely met. There is nothing to suggest that the evaluator will engage in any unethical conduct, but neither is there information to suggest that the evaluator has considered all of the ethical problems that can arise during an evaluation study.

One way in which the evaluator has been responsive to potential ethical problems is by requiring that evaluation reports will be approved by a panel of educators before release (page 6). This panel will provide guidance on several ethical issues. However, the evaluator has not considered the two other issues treated by this criterion. The evaluator should provide evidence that he intends to comply with protection of human subjects guidelines as applicable in the study. Also, the evaluator should guarantee that the data collected during the study will not be released to unauthorized personnel or be used inappropriately.

B. Protocol: Are appropriate protocol steps planned?				
1. Are appropriate persons contacted in the appropriate sequence?	(Yes)	No	?	NA
2. Are Department policies and procedures to be followed.	(Yes)	No	?	NA

The evaluator has given adequate consideration to Protocol criterion in the design. In this case, the evaluator plans to clear virtually everything through the chief school officers (page 11). Although more specific protocol steps will evolve during the evaluation study, the evaluator has set a procedure to meet initial protocol needs.

Summary

As was noted earlier, the fictitious evaluation design of the Hartman Reading Program is neither all good nor all bad. The design has both strengths and weaknesses, and use of the checklist has helped identify them. However, simply using the checklist is not enough. Information about the evaluation design from the checklist should be provided to the evaluator so that weaknesses in the design can be discussed and corrected before the evaluation begins. By so doing, an important step toward producing a helpful evaluation study will have been taken.

IV. A REVIEW OF PREVIOUS WORK AS A BASIS FOR DETERMINING THE ADEQUACY OF AN EVALUATION DESIGN

Most educators who have ever been involved in evaluation have worried about determining the quality of the evaluation effort. Although implicit standards have long been used in determining the quality of evaluation plans, evaluation specialists have only recently begun to develop an explicit, well defined basis for determining the adequacy of such designs.

Michael Scriven (1969) first coined the term "meta-evaluation" to refer to the evaluation of evaluation. Since then, several evaluators have proposed standards for determining the quality of evaluation designs.

Many specialists' proposed standards have evolved from their training backgrounds or from definitions of evaluation that they have adopted. Consideration of such proposals can help one understand the evolution of the checklist offered in the previous section. Because of the considerable effort that has recently gone into the development of a basis for evaluating evaluation designs, it is important to draw as much usable information as possible from these efforts.

Bases for judging evaluation designs have generally been presented in one of three ways: (1) as guidelines that provide a format for evaluation designs, (2) as essays describing elements of a good evaluation, or (3) as checklists that guide the application of standards to evaluation designs. Examples of each are included in this section.

Guidelines for Evaluation Designs

Worthen and Sanders (1973) suggested the following format for evaluation designs, a set of elements that could be considered to all evaluation designs.

SUGGESTED FORMAT FOR EVALUATION PROPOSALS [3]

I. Rationale (Why is this evaluation being done?)

II. Objectives of the Evaluation Study

- A. What will be the product(s) of the evaluation study?
- B. What audiences will be served by the evaluation study?

III. Description of the Program Being Evaluated

- A. Philosophy behind the program
- B. Content of the program
- C. Objectives of the program, implicit and explicit
- D. Program procedures (e.g., strategies, media)
- E. Students
- F. Community (federal, state, local) and instructional context of program

IV. Evaluation Design

- A. Constraints on evaluation design
- B. General organizational plan (or model for program evaluation)
- C. Evaluative questions
- D. Information required to answer the questions
- E. Sources of information; methods for collecting information
- F. Data collection schedule
- G. Techniques for analysis of collected information
- H. Standards; bases for judging quality
- I. Reporting procedures
- J. Proposed budget

V. Description of Final Report

- A. Outline of report(s) to be produced by evaluator
- B. Usefulness of the products of the study
- C. Conscious biases of evaluator that may be inadvertently injected into the final report

---

[3] Worthen, B. R. and Sanders, J. R. Educational Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones, 1973. p 301.

A similar format was suggested by Stake (1969) in the following guide for a final evaluation report: [4]

Section I - Objectives of the Evaluation

A. Audiences to be served by the evaluation
B. Decisions about the program, anticipated
C. Rationale, bias of evaluators

Section II - Specification of the Program

A. Educational philosophy behind the program
B. Subject matter
C. Learning objectives, staff aims
D. Instructional procedures, tactics, media
E. Students
F. Instructional and community setting
G. Standards, bases for judging quality

Section III - Program Outcomes

A. Opportunities, experiences provided
B. Student gains and losses
C. Side effects and bonuses
D. Costs of all kinds

Section IV - Relationships and Indicators

A. Congruences, real and intended
B. Contingencies, causes and effects
C. Trend lines, indicators, comparisons

Section V - Judgments of Worth

A. Value of outcomes
B. Relevance of objectives to needs
C. Usefulness of evalution information gathered

[4] Stake, R. E. Evaluation design, instrumentation, data collection, and analysis of data. Educational Evaluation. Columbus, Ohio: State Superintendent of Public Instruction, 1969.

Essays About Evaluation Quality

Essays on educational evaluation offer general statements about the elements of good evaluation, and provide a second source of standards. One such essay, by Worthen (1973), "A Look at the Mosaic of Educational Evaluation and Accountability," covered the following considerations:

1. Conceptual Clarity

   Conceptual clarity is an essential feature of any good evaluation plan. By "conceptual clarity" I refer to the evaluator's exhibiting a clear understanding of the particular evaluation he is proposing. Is he planning a formative or summative evaluation? Is it a comparative evaluation design or a single program evaluation? Is the evaluation to be goal-directed, with the design built around the measurement of attainment of specific objectives, or goal-free with the design built around lists of evaluative questions generated independently of the goals? Answers to these questions should be apparent in any good evaluation plan; for without clarity on these points, proper evaluation could occur only by chance.

2. Characterization of Program

   No evaluation is complete without a thorough, detailed description of the program or phenomenon being evaluated. Without such characterization, judgments may be drawn about a program which never really existed. For example, the concept of team teaching has fared poorly in several evaluations, resulting in a general impression that team teaching is ineffective. Closer inspection shows that the methods frequently labeled "team teaching" provide almost no real opportunities for staffs to plan together or work together in direct instruction. Obviously, a better description of the phenomenon would have avoided these misinterpretations completely. One simply cannot evaluate adequately that which he cannot describe accurately.

3. Recognition and Representation of Legitimate Audiences

   Any evaluation will be adequate only to the extent to which it provides for obtaining input from and reporting to all legitimate evaluation audiences. An evaluation of a school program which answers only the questions of the school staff and ignores questions of parents, children and community groups is inadequate. Each legitimate audience must be identified and the objectives or evaluative questions of that audience considered in designing a plan for data collection. Obviously,

some audiences will be more significant than others and some weighting of their input might be necessary. Correspondingly, the evaluation plan should provide for receipt of appropriate evaluation information by each audience which has a potential interest in the program.

4. Sensitivity to Political Problems in Evaluation

Many a good evaluation, unimpeachable in all of its technical details, has failed because of its political naivete. It is pointless to promise to collect sensitive data--e.g., principals' ratings of teachers--without first obtaining permission from the office or individual who controls those data. Procedures governing access to data and data sources, and safeguards against misuse of evaluation data must be agreed upon early in the project. Steps must be taken to guarantee that program staff have opportunities to correct factual errors in evaluation reports without compromising the evaluation itself. These issues exist in almost every evaluation and the more explicitly they are dealt with, the more likely the evaluation is to survive political pressures.

5. Specification of Information Needs and Sources

Good evaluators tend to develop and follow a blueprint which tells them precisely what information they must collect and through what sources that information is available. At the very least, they know how (as Scriven puts it) to lay snares at critical points in the game trails. Conversely, the novice evaluator goes about randomly turning over stones or beating the brush to see what he can find. No evaluation can depend on a random, scattered "here a little, there a little" approach to collecting data. An adequate evaluation plan specifies at the outset the information which must be collected. If the evaluation is goal-directed, the plan will specify information that will help to determine whether the objectives were attained. If the evaluation is built around evaluative questions (of the "What would you need to know to decide whether the program was a success or a failure?" variety), the evaluation plan should specify information which, when collected, will answer those questions. And in every case, specifying needed information leads logically to identification of the sources from which that information can be obtained. Failure to attend to these seemingly pedestrian but truly critical steps is one of the greatest single reasons that many evaluations produce little useful information.

6. Comprehensiveness/Inclusiveness

This category is really an elaboration of the previous one. No evaluation can hope to collect all of the relevant data--nor would it be desirable to do so, since there will always be inconsequential and trivial data not worth the bother to collect. Collecting too much data is seldom the concern, however. The greater problem is collecting enough data--or more precisely, collecting data on enough important variables to be certain one has included in the evaluation all the major considerations which are relevant. A good evaluation includes all of the main effects, but also includes provisions for remaining alert to unanticipated side effects. A good comparative evaluation doesn't stop with comparing the experimental arithmetic program with a control group which receives no arithmetic instruction. It goes on to identify the critical competitors--SMSG math, Cuisennaire Rods, and so forth--and compares their new program with those for which costs are roughly comparable. In short, the weak evaluation is almost always characterized by a narrow range of variables and omission of several important variables. The wider the range and the more important the variables included in the evaluation, the better it generally is.

7. Technical Adequacy

More evaluations founder on this shoal than on almost any other, and this is due to the scarcity of educational evaluators who are even marginally competent in technical areas. Good evaluations are dependent on construction or selection of adequate instruments, the development of adequate sampling plans, and the correct choice and application of techniques for data reduction and analysis. Volumes have been written on educational measurement, sampling, and statistics and it would be pointless to try to review that knowledge here. Suffice it to say that competence in these areas is essential to most evaluations. Without knowledge and control of these tools of his trade, the evaluator has little hope of producing evaluation information which meets scientific criteria of validity, reliability and objectivity.

8. Consideration of Program Costs

Educators are not econometricians and should not be expected to be skilled in identifying all the financial, human or time costs associated with programs they operate. That bit of leniency cannot be extended to the evaluator, however, for it is his job to bring these factors to the attention of teachers and administrators who are responsible for the programs. Educators are often faulted for choosing the more expensive

of two equally effective programs, just because the expensive one is packaged more attractively or has been more widely advertised. The real fault lies with the evaluations of those programs which fail to focus on cost factors as well as on other variables. As any insightful administrator knows, costs are not irrelevant, and it is important for him to know how much program X will accomplish and at what cost so he may know what he is gaining or giving up in looking at other options which vary in both cost and effectiveness.

9. Explicit Standards/Criteria

It is always a bit disconcerting to me to read through an evaluation report and be unable to find anywhere a statement of the criteria or standards which were used to determine the program's success or failure. The measurements and observations taken in an evaluation cannot be translated into judgments of worth without standards or criteria. Is an in-service program for teachers successful if 75% of the teachers attend 75% of the meetings? That all depends on the standard that is set for the program. What about a 60% attendance rate in a high school English class--is that good or bad? Again it depends on the standard. If it is a regular English class, with a standard of 95%, 60% looks pretty bad. But in an English class for rehabilitated dropouts who work part-time to support their parents, the standard might be 50% and the attendance rate of 60% might be quite acceptable. Every good evaluation will include a statement of standards and criteria.

10. Judgments and/or Recommendations

The only reason for insisting on explicit standards or criteria is that they are the stuff of which judgments and recommendations are made, and these judgments and recommendations are the sine qua non of evaluation. An evaluator's responsibility does not end with the collection, analysis, and reporting of data. The data do not speak for themselves. The evaluator who knows those data well is in the best position to apply standards for judging effectiveness. Making judgments and recommendations is an essential part of the evaluator's job. An evaluation without judgments is as much an indictment of its author's sophistication as one with recommendations that are not based on the data.

11. Reports Tailored to Audiences

I argued a few minutes ago that there are multiple audiences for most evaluations and these audiences have different informational needs. For example, when you complete an evaluation, your colleagues in evaluation will be interested in a complete, detailed report of your data collection procedures, analysis techniques, and the like. Not so for the school board, or the PTA or the little old lady in tennis sneakers who heads the local taxpayer group. These audiences do not share the evaluator's grasp of technical details or his interest in test reliability and validity or the appropriate choice of an error team in a randomized blocks design. The evaluator will have to tailor reports for these groups so that they depend on non-technical language, and he must avoid over-use of tabular presentation of data analyses. A typical evaluation might produce one omnibus technical evaluation report which self-consciously includes all the details and one or more non-technical evaluation report(s) aimed at the important audience(s).

Another notion should be inserted here as well--that of interim or even continual reporting of evaluation findings. Timeliness is an important concern in evaluation. Information that is presented too late to affect the decision for which it is relevant is useless. Good evaluations will not depend solely on the printed word, but will include a variety of report formats--including "hot-line" telephone reporting--so the information is reported whenever it is needed to make a particular decision.

Other general standards which have been widely used include the following, developed by Stufflebeam et al. (1971)[5]:

1. Internal validity. Does the evaluation design provide the information it is intended to provide? The results of the evaluation study should present an accurate and unequivocal representation of the object being evaluated.

2. External validity. To what extent are the results of the study generalizable across time, geographical environment and human involvement? In many small evaluation studies, the concept of external validity is irrelevant since the evaluator is interested in collecting and interpreting information about one specific program at one point in time. However, the concept may be quite important in large-scale evaluation studies where sampling is used and findings must be generalized back to the total population.

---

[5] Stufflebeam, D. L. et al. Educational Evaluation and Decision-Making in Education. Itasca, Illinois: Peacock, 1971.

3. Reliability. How accurate and consistent is the information that is collected? The evaluator should be quite concerned about the adequacy of his measures since his results can only be as good as the information on which they are based.

4. Objectivity. How public is the information collected by the evaluator? The evaluator should strive to collect information and make judgments in such a way that the same interpretations and judgments would be made by any intelligent, rational person evaluating the program.

5. Relevance. How closely do the data relate to the objectives of the evaluation study? Defining objectives for an evaluation study enables the evaluator to check himself on the relevance of his activities.

6. Importance. Given a set of constraints on the design of an evaluation study, what priorities are placed on the information to be collected or program components to be evaluated? It is often tempting to study one relevant aspect of a program in depth and to collect much information which may subsequently prove to be less important at the conclusion of the study than less detailed information about another aspect might have been. It is the responsibility of the evaluator to set priorities on the data to be collected.

7. Scope. How comprehensive is the design of the evaluation study? There are a wide variety of considerations to explore, as emphasized in several papers presented in the previous chapter. The evaluator must consciously avoid the possibility of developing "tunnel vision" by taking a wholistic approach to program evaluation.

8. Credibility. Is the evaluator believed by his audiences? Are his audiences predisposed to act on his recommendations? The evaluator-client relationship is an important one if the evaluator wants his efforts to have some impact on the program he is evaluating.

9. Timeliness. Will evaluation reports be available when they are needed? Many evalutors have missed the chance to influence action because they reported too much, too late. When decisions affecting a program are being made, any reliable information is better than none. The provision of interim, often informal, reports will help to avoid this problem of being too late to influence the decision.

10. Pervasiveness. How widely are the results of the evaluation study disseminated? It is true that, in many cases, only one audience needs to be addressed. However, the evaluator is responsible to provide the results of his study to all individuals or groups who should know about the results.

11. Efficiency. What are the cost/benefits of the study? Have resources been wasted when that waste could have been avoided? Operating under the constraints imposed on most evaluation studies, the evaluator is responsible for making the best possible use of material and human resources available to him.

Checklists That Guide the Application of Standards to Evaluation Designs

Checklists which guide the application of standards to evaluation designs or reports are a third source of standards. These checklists cover many general concerns; the most useful checklists also include highly specific, comprehensive standards which can assist in determining the quality and completeness of evaluation designs.

Each existing checklist seems unique in form, content and purpose; nevertheless, many share common characteristics. Generally, checklists for judging evaluation designs include considerations of the scientific or technical adequacy of the evaluation, the practicality and cost efficiency of the design, the usefulness of the data to be collected, and the responsiveness of the design to legal and ethical issues.

Four checklists for judging evaluation designs are described below. The first of the checklists, that written by Stake (1970), contains five general areas in which evaluation designs are to be judged: (1) the evaluation itself, (2) specifications of the program being evaluated, (3) program outcomes, (4) relationships and indicators, and (5) the program's overall worth.[6] Each

[6] Stake, R. E., A Checklist for Rating an Evaluation Report, Unpublished manuscript, October, 1970.

general area, in turn, covers specific considerations which, when relevant, are to be judged on their individual adequacy.

The checklist by Bracht (1973) includes six areas on which evaluation designs should be judged: (1) communication, (2) importance of the evaluation, (3) design for making judgments, (4) design for obtaining descriptive data, (5) reports, and (6) concerns.[7] Detailed questions are included within each of these six areas of concern.

Stufflebeam's (1974) checklist covers six aspects of the design: (1) conceptualization of the evaluation, (2) socio-political factors, (3) contractual/legal arrangements, (4) the technical design, (5) the management plan, and (6) moral/ethical/utility questions.[8] Rather than questioning the adequacy of certain aspects of evaluation design, Stufflebeam seeks specific information that should be included in an evaluation design.

The final checklist, compiled by Smith and Murray (1974), includes a number of questions from other checklists.[9] Smith and Murray address three areas of evaluation design: (1) content descriptions, (2) evaluation activities/results, and (3) document characteristics. Each of these major areas is further divided into two subareas with appropriate exemplary questions designed to determine the adequacy of those subareas.

Guidelines for evaluating school practices provide another source of evaluation design standards. Directions for program audits produced by the

---

[7] Bracht, G. H., Evaluation of the Evaluation Proposal, Unpublished manuscript, 1973.

[8] Stufflebeam, D. L., An Administrative Checklist for Reviewing Evaluation Plans, Unpublished manuscript, April 1974.

[9] Smith, N. L., and Murray, S. J., Evaluation Review Checklist, Unpublished manuscript, 1974.

federal government and directions for evaluation audits produced by auditing agencies contain examples of such criteria. Such guidelines are also available from the National Study of School Evaluation (NSSE) Evaluative Criteria[10] for secondary schools, middle schools, elementary schools and multicultural programs. These guidelines, used by accreditation teams throughout the country in evaluating school programs, contain a comprehensive list of school characteristics useful in checking the completeness of a design for evaluating a school program.

Summary

The review provided in this section demonstrates the extensiveness of the work that has been done by educators in producing criteria for judging evaluation designs and reports. Because of this considerable effort, the practice of judging evaluation designs and reports is becoming more and more common among educators who are involved with producing or using evaluation studies on a daily basis. And, while there are many differences among the various sets of criteria presented in this section, many common threads of thought can be found. The criteria presented earlier in this paper reflect those common elements.

[10] Evaluative Criteria (Fourth Edition), National Study of Secondary School Evaluation, Washington, D. C., 1969.

References

Astin, A. W., and Panos, R. J. The evaluation of education programs. In R. L. Thorndike (Ed.) Educational measurement. Washington, D. C.: American Council on Education, 1971.

Bracht, G. H. Evaluation of the evaluation proposal. Unpublished manuscript, 1974.

National Study of Secondary School Evaluation. Evaluative Criteria (4th ed.). Washington, D. C.: Author, 1969.

Scriven, M. An introduction to meta-evaluation. Educational Product Report, 1969, 2, 36-38.

Smith, N. L., and Murray, S. J. Evaluation review checklist. Unpublished manuscript, 1974.

Stake, R. E. Evaluation design, instrumentation, data collection, and analysis of data. Educational Evaluation. Columbus, Ohio: State Superintendent of Public Instruction, 1969.

Stake, R. E. A checklist for rating an evaluation report. Unpublished manuscript, 1970.

Stufflebeam, D. L., et al. Educational evaluation and decision-making in education. Itasca, Illinois: Peacock, 1971.

Stufflebeam, D. L. An administrative checklist for reviewing evaluation plans. Unpublished manuscript, 1974.

Worthen, B. R., and Sanders, J. R. Educational evaluation: theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Worthen, B. R. A look at the mosaic of educational evaluation and accountability. Research, Evaluation, and Development Paper Series. Portland, Oregon: Northwest Regional Educational Laboratory.

Wright, W. J. and Worthen, B. R. Standards and procedures for development and implementation of an evaluation contract. Alaska Department of Education, 1975.